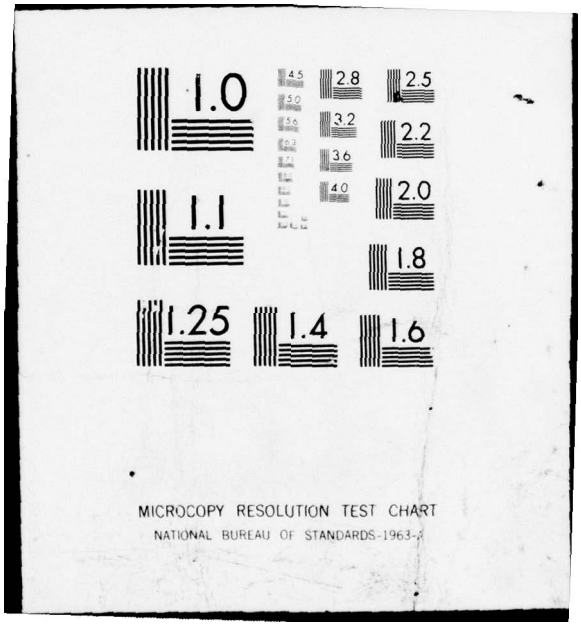
AD-A044 713 STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE F/G 6/4
THREE REVIEWS OF J. WEIZENBAUM'S COMPUTER POWER AND HUMAN REASO--ETC(U)
NOV 76 B G BUCHANAN, J LEDERBERG, J MCCARTHY MDA903-76-C-0206
UNCLASSIFIED STAN-CS-76-577 NL

| OF |
AD
A044 713

END
DATE
FILMED
10-77
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Stanford Artificial Intelligence Laboratory
Memo AIM-291

November 1976

Computer Science Department
Report No. STAN-CS-76-577

Three Reviews of J. Weizenbaum's

COMPUTER POWER AND HUMAN REASON

by

Bruce G. Buchanan
Joshua Lederberg
John McCarthy

Research sponsored by

Advanced Research Projects Agency
ARPA Order No. 2494

COMPUTER SCIENCE DEPARTMENT
Stanford University

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: STAN-CS-76-577, AIM-291	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Three Reviews of J. Weizenbaum's COMPUTER POWER AND HUMAN REASON.	5. TYPE OF REPORT & PERIOD COVERED: Technical rept.
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Bruce G. Buchanan, Joshua Lederberg and John McCarthy	8. CONTRACT OR GRANT NUMBER(s): MDA903-76-C-0206, DAHC15-73-C-0435
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Artificial Intelligence Laboratory, Stanford University, Stanford, California 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ARPA Order 2494
11. CONTROLLING OFFICE NAME AND ADDRESS: Col. Dave Russell, Dep. Dir., ARPA IPT, ARPA Headquarters, 1400 Wilson Blvd. Arlington, Virginia 22209	12. REPORT DATE: November 1976
	13. NUMBER OF PAGES: 28 pages
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Philip Surra, ONR Representative, Durand Aeronautics Building Room 165, Stanford University, Stanford, California 94305	15. SECURITY CLASS. (of this report): 15
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Releasable without limitations on dissemination.

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three reviews of J. Weizenbaum's Computer Power and Human Reason (W. H. Freeman and Co., San Francisco, 1976) are reprinted from other sources. A reply by Weizenbaum to McCarthy's review is also reprinted.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

094120 4B

Stanford Artificial Intelligence Laboratory
Memo AIM-291

November 1976

Computer Science Department
Report No. STAN-CS-76-577

Three Reviews of J. Weizenbaum's

COMPUTER POWER AND HUMAN REASON

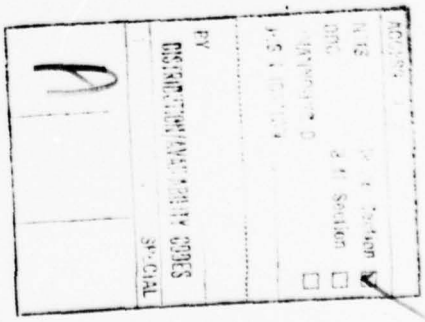
by

Bruce G. Buchanan
Joshua Lederberg
John McCarthy

ABSTRACT

Three reviews of Joseph Weizenbaum's Computer Power and Human Reason (W.H. Freeman and Co., San Francisco, 1976) are reprinted from other sources. A reply by Weizenbaum to McCarthy's review is also reprinted.

The work reported here was funded by the Advanced Research Projects Agency, under ARPA Contracts DAHC 15-73-C-0435 and MDA 903-76-C-0206.



Review of Joseph Weizenbaum's **Computer Power and Human Reason**

by
Bruce G. Buchanan
Adjunct Professor of Computer Science
Stanford University

The following review is to appear in PHAROS, Fall 1976.

Weizenbaum's **Computer Power and Human Reason** is a book about the immorality of applying computers to tasks that ought to be done only by humans. The author speaks out against improper and immoral uses of the tools available to us, especially computers. Some readers will find his description of computers illuminating and alarming, the rest of us can still benefit from the challenge to sharpen our moral criteria.

The author is a leading computer scientist teaching at M.I.T. who is disturbed by both the strongly pro-technology, and the anti-humanist statements of some computer scientists. He is also concerned about the public's passive acceptance of technologists' definitions of social problems and their subsequent technological solutions. Examples supporting both concerns are numerous, and readers need to remind themselves often that there is a point to the examples, many of which are very pointed criticisms of individuals. Much (perhaps too much) of the book is devoted to reasons why the book was written.

The last chapter (10) contains the material most worth reading, with the first chapter providing a very readable introduction. With some misgivings at trying to collapse 250 pages into a few lines, we can summarize the main theme of the book in an informal argument:

(P1) It is wrong for us to cause people to be treated as less than "whole persons". (e.g., "An individual is dehumanized whenever he is treated as less than a whole person." p. 266).

(P2) Computers can never fully understand human problems [because they must ignore aspects of human experience that are not describable in a language].

(P3) Therefore, it is wrong for us to command computers to deal with human affairs, i.e., to perform tasks requiring understanding of and empathy for human problems [esentially because computers lack a person's complete view of other persons].

Proposition P1 is a presupposition for the entire book.

But the concepts of the whole man or of man as object are not as simple and obvious as the author would have us believe, and the moral dictum in proposition P1 is not easy to apply, for in some circumstances it is desirable to treat persons as anonymous entities, as in voting. Much of the material in Chapters 4-9 is an exposition of a non-mechanistic view of man, in support of proposition P2. Chapters 2-3 discuss the abilities of computers and how they work. Intertwined throughout all the chapters is evidence that people want to and can apply computers in human affairs as well as evidence that it leads to evils. Now the problem is understanding these propositions and deciding whether one believes them to be true. In the book, this is largely left as an exercise for the reader.

Any useful tool can always be misused; the particular dangers of computers stem from their versatility and complexity. Because computers are general symbol manipulation devices, they are versatile enough to be used, and misused, in applications that affect life or that have serious, irreversible social consequences. And computer programs are often so complex that no person understands the basis for the program's decisions. As a result, we are told, managers, physicians, and decision makers of all sorts do not feel responsible for the consequences of decisions made by computer programs. A large portion of the book is an elaboration on these dangers. These dangers are also used to support proposition P3, above: that there are decision making tasks we ought not turn over to computers.

The computer programs that most concern the author are the complex problem solving programs developed under the general name of artificial intelligence. These are classified as (a) simulations of human problem solving (cognitive simulations), (b) programs that solve difficult problems without trying to duplicate human methods (performance programs), and (c) programs, or other work, that explore theoretical issues in computing (these he ignores). The claims made about cognitive simulations and performance programs seem to disturb Weizenbaum more than the work itself. Perhaps those claims, made ten and twenty years ago, should have been more cautiously phrased to say, for example, that computers will successfully simulate some [not all] aspects of human thought and complement [not replace] human problem solvers. Several examples are taken from psychiatry, partly because others read too much into a simple program he wrote ten years ago that mimics some conversational aspects of a therapist, a fact that profoundly disturbs him.

In any case, Weizenbaum's arguments are aimed at showing that computer capabilities are not coextensive with human capabilities. If the arguments are correct, then computer programs cannot successfully duplicate all aspects of human intelligence (proposition P2, above). Whether they could sometimes be used, morally and profitably, in place of humans still depends on understanding the concept of dehumanization in proposition P1.

This viewpoint can be pushed to extremes to argue against any use of computers, in which case even the worst human decision makers are seen as better qualified to make social decisions than the best computer programs ever will be. What seems more credible is that computer programs are more dangerous in the hands of poor managers than in competent hands. And Weizenbaum obliquely gives us a criterion for competence: the competent manager understands the basis for the program's decisions and maintains an ability and willingness to override the programs. On the other side of the coin, a program becomes less dangerous (in any hands) if it can demonstrate its line of reasoning from problem description to problem solution, can be queried about its assumptions and methods, and otherwise opens itself to understanding.

Admittedly, computer programs are not easily understood (the main point of chapter 9). This is a great shortcoming, but not one that has escaped notice. Also, managers (of many sorts) have admittedly abdicated some responsibility to technology. In just this sense, physicians are chided for becoming "mere conduits between their patients and the major drug manufacturers" (p.259). But competent managers and physicians will first understand the scope and limitations of their tools before using them.

Weizenbaum is asking us not to do research on programs, methods or tools with obvious potential gross misuses: he finds no benefits that are worth the price of meddling with tools "that represent an attack on life itself" or that substitute for "interpersonal respect, understanding, and love" (p. 269). Incidentally, this is the same theme as expressed in his letter to **Science** attacking recombinant DNA research (1). He also advocates renunciation of projects with "irreversible and not entirely foreseeable side effects."

The guidelines that he gives are certainly incomplete for research on energy, communication, transportation -- and almost anything interesting enough to be applied in the next century -- would have unforseeable side effects or could be used to assault life. They are offered as expressions of his own subjective criteria, and perhaps because they are subjective they cannot be expressed adequately in the language of the brain's left hemisphere (as he reminds us in another context). Such guidelines, even when precise, also fail to admit the value of research aimed at defining the limits of what computers can do by working on programs at the boundaries between men and machines.

He says he is not asking others to adopt his own criteria, that the book advocates exercising our own courage to say NO when our "inner voice" tells us an act is wrong (p. 276). Since we do not know what another person's conscience tells him, however, we do not necessarily feel safer knowing it has been consulted.

---

(1) **Science** , July 2, 1976, p.6.

Weizenbaum's attacks on his colleagues seem to reduce to the question whether or not they have listened to their own inner voices.

The same concerns were raised in less inflamatory books and articles by Norbert Wiener (2). Wiener's solution, if it may be called that, is not to stop work on some research but to

> "render unto man the things which are man's and unto the computer the things which are the computer's. This would seem the intelligent policy to adopt when we employ men and computers together in common undertakings. It is a policy as far removed from that of the gadget worshiper as it is from the man who sees only blasphemy and the degradation of man in the use of any mechanical adjuvants whatever to thoughts."

This view of the symbiotic relationship of men and machines is a much more constructive one than Weizenbaum's. It places the computer in the hands of problem solvers, and not the other way around. In this view, there are still interesting questions for computer specialists: how can a program provide a manager (problem solver, decision maker, etc.) with enough information that he can accept responsibility for its output? How can the program convey its scope and limitations to the manager? How can we design programs that are more useful for managers -- i.e., easier and more pleasant to use, easier to understand, more knowledgeable and more flexible?

Another recent book dealing with the relationship of man and machine is **Zen and the Art of Motorcycle Maintenance** (3). It too builds on the premise that scientific, logical inquiry leads to only one kind of truth, that the subjective, intuitive, emotional side is necessary for human interaction and is equally legitimate. There is a similarity in the message, but a world of difference in style: while Pirsig's novel evokes emotions, Weizenbaum prescribes them.

A distressing undercurrent through the whole book is its anti-rationalism. In discussing dangerous research in the book, with recombinant DNA used as an example, he questions the need to give any justification for stopping the research. For example,

---

(2) E.g., N. Wiener, **God & Golem, Inc.** (M.I.T. Press, Cambridge, 1964).

(3) R. Pirsig, **Zen and the Art of Motorcycle Maintenance** (Morrow, New York, 1974).

> "Is not the overriding obligation on men ... to exempt life itself from the madness of treating everything as an object, a sufficient reason, and one that does not even have to be spoken? Why does it have to be explained? It would appear that even the noblest acts of the most well-meaning people are poisoned by the corrosive climate of values of our time." (pp.260-61).

But this irrationalistic sentiment ignores the value of giving reasons for halting research on projects with great potential benefits for human health as well as dangers. If scientists failed to provide reasons for their decisions, either to halt or continue lines of research, how can we ever expect informed decisions from the public and legislative representatives? If there is madness in treating people "as objects" there is just as much madness in assuming that one's own research decisions require no justification.

In summary, the main issues of the book are important for everyone, and are especially directed at persons working with computers. In spite of the backbiting, the digressions, and the vague language in which the perceived evils are (and perhaps must be) described, the book deserves discussion and contemplation. It is not the kind of book that can be taken literally but it does raise questions that all scientists need to answer for themselves.

Review of Joseph Weizenbaum's Computer Power and Human Reason

by
Joshua Lederberg

Professor of Genetics
Stanford University

The following review of J. Weizenbaum's book was solicited and accepted by the N.Y.Times Book Review 1 March 1976 but to the best of my knowledge has not appeared in print. J. Lederberg

"Computer power and human reason" is a mosaic of well-reasoned analysis and passionate pleading on the nature of computers and of man, and about the place that computers (read "technology" if you wish) should have in human affairs. Prof. Weizenbaum is particularly exercised about the claims for and prospects of AI, "artificial intelligence", the efforts to emulate and bolster human reasoning processes as programs in computers. A well known computer scientist at M.I.T, who has made significant contributions to AI, he writes that he was moved to write this tract when people responded to one of his programs as if it were an empathic companion .Tnis over-estimation of, and over-dependence upon computers,he believes to be both symptom and cause of global predicaments with more horrors to come.

Weizenbaum may still be too much a technocrat: witness his oversimplifications of social movements as the immediate fruits of technical innovations. (There is more to the history of the internal migration and urbanization of American blacks than the introduction of the mechanical cotton-picker in the 1950's.)But to dwell on these would do too little justice to the other fundamental issues that Weizenbaum raises.Indeed he might be the first to deplore his own vestigial technocratic biases, when they are inconsistent with his fundamental ethical philosophy. Nor should one harp on his ad hominem attacks on some of his colleagues, his bludgeoning them with selected quotations from writings of 20 years ago,which I suspect he will view as a lapse more from enthusiasm than from malicious intention.

Most readers will follow the author's advice to skip over the early chapters which detail the fundamental logic of the computer -- these would make another, peerless book for explaining Turing's work on the fundamental logic of computing machines to the lay reader. The basic philosophical and policy issues do not need this detail, and are best scrutinized by reading the book back-to-front: few readers with the interest and general intellectual grounding to digest this work

critically, will need to be reintroduced to the fundamentals. Others will be attracted by pages of lyrical anti-technology slogans, which the author's technical reputation will make the more persuasive. Since the author categorically rejects "instrumental reason" in its application to human affairs, it is difficult to engage him in a discussion of his particular policy concerns.

Weizenbaum makes a conscientious effort to distinguish his assertions of faith from the scientific consensus; but the non-specialist reader will still have to look closely to be sure. Perhaps in fields like the physiology of the right versus left brain he has already persuaded himself that contemporary speculations are proven realities about the location of human rational functions. But others should be cautioned that we still know even less about the organization of human intellect than Weizenbaum stipulates.

During the early adolescence of computer science in the early 50's, many workers made extravagant prophecies about the ease with which the new machines would be programmed to match human problem-solving behavior -"within the visible future", we were told, machines would conduct mechanical translations of high quality (especially from Russian into English). They would play chess to the disadvantage of the masters, and they might then be ready to take over many of the higher-level functions of management in industry, and of command and control in the military. Within a few years, the power of machines to manipulate bits of information had been enhanced a million-fold: what more could one ask as the basis for these new powers?. It is no surprise, and by now no news, that these prophecies were simply wrong; and the wiser among us should have learned not to make technological forecasts where we simply had new tools, but no real insight into the structure of the tasks they were to address. Surely, as Weizenbaum insists, there are few things less well understood than human creative imagination. His own prophecy is that this will NEVER be emulated to any significant measure by computing machines. This hypothesis is beyond the range of scientific criticism, short of tangible advances too much to hope for right away; but his arguments are mainly repetitious assertions of his personal faith.

No, there is one more persuasive kernel: namely that the world-knowledge which underlies human understanding (compassion and judgment) needs the life-long experience of having been human -- in a word, of having shared love .It is unlikely and undesirable that machines be offered that privilege; then many realms will be uniquely human. Indeed we must make equally sure that the fellow-creatures to whom we confide our trust for ethical and esthetic leadership justify this on the same grounds. The abrogation of human responsibility for moral decision whether it be out of lazy delegation to machines, or superstitious deference to super-human abstractions can indeed once again ignite the holocaust.

Weizenbaum's pleading overreaches this sufficient argument to an out-and-out obscurantism about the fundamental non-comprehensibility of the human brain, which adds little to the debates between

vitalists and mechanists of the last two centuries. It is a sterile debate; and scientists can contribute more by trying to find what can be learned about our own nature, and putting it to human good, than arguing what may or may not be ultimately knowable. As with his concern about the bounds of AI, the mischief of such criticism is that it may disparage the work of investigators with more concrete, modest and achievable goals. The view that the core of the cell's reproductive capability was unknowable in chemical terms bears much of the onus for long delays in our understanding of the structure of DNA. After the fact, this proved to be remarkably simple.

While we should not offer love to the machine, there is much to be said for permitting it to evolve, that is to nurture the growth of more and more complex programs. These are initiated by human intelligence, but grow from the dynamics built into the starting program itself. It is hard to see how some of the more complex problems to be addressed can be solved by programs that are explicitly written in detail by human authors. Then I agree with Weizenbaum that we can longer claim to have a full-fledged explanation of a phenomenon, merely through having generated a model for it. We may even have substantial power to solve problems without necessarily "understanding" them. (Unfortunately, this criticism does little to help us recognize true understanding by any objective criterion). Furthermore, we should not trust such complex programs merely because we believe we were sufficiently intelligent in our original design plans. Instead, the program will have become another experiment, to be validated only by experience. Much the same ought to be said for other areas of human aspiration, like politics.

Weizenbaum is particularly critical of the use of the computer in the role of psychotherapy -- doubtless in consternation that the machine's patrons believed they were talking to a sympathetic, understanding 'person'. This criticism raises a number of issues that deserve more analytical attention: 1) Is it true that the patrons were confused, or do many of them find some service in the 'dialogue' well knowing that they are at best talking to themselves? and 2) Can a therapeutic utility by this modality be empirically validated and economically justified? At the moment our answers to these questions are speculative and anecdotal, and I would not substitute my own critical skepticism about this approach for a conclusive dismissal of it. Weizenbaum's criticism indeed may misapprehend the role of psychotherapy as a source of self-insight -- where the patient himself must do most of the work at achieving human understanding -- and machines may well be expected to play some useful role in this process (no differently than, say, the reading of a book -- and in a similar analogy to computer-assisted-instruction in other domains.)

Throughout his book, Weizenbaum oscillates between a disparagement of the potential and actual accomplishments of AI, dismay at what he sees as excessive faith and dependence on this technology, and concern for some potential abuses of its development, should it be realized. That policy-makers, the public, and computer scientists alike should take a more critical and pragmatic view of the field than the zealots of 20 years ago may be granted; many well-informed people within the field clearly do, without having reacted as strongly as Weizenbaum.

The abuses might be either ideological or technological. If human intelligence were more successfully mirrored in the machine, will that not justify treating human beings as if they were MERE machines? His position on this issue is colored by the experience of Nazi Germany; but the argument is confused. The most savage tyrannies that I can find in history, including Nazism, had no doubt about a unique elan-vital -- just that one folk or credo had more than an equal share. People who are philosophically concerned about the mechanistic basis of life are also overawed by its complexity, and too concerned about learning more about it to occupy themselves with holy wars. They are the least likely to be sacrificing either people or machines on the grounds of ideological conviction.

The historical record is less reassuring about the augmentation of power in the hands of irrational man: we can still argue about the case against Prometheus,Gutenberg, Galileo,or Faraday -- not to mention Oppenheimer -- but by that very token, I do not share Weizenbaum's confidence in deciding which innovations are dangerous. He points to very real concerns about machines that could interpret speech (while denying their feasibility). Yes-- they might make large scale wire-tapping irresistible, and perhaps undo the virtues of the telephone as a medium of private communication. They might also relieve millions of office-persons from the mindless tasks of transcribing the words of others, and free them for more creative responsibilities. Both of these contingencies lay heavy burdens on the adaptability of our social institutions, and it is important that we be alerted to them.

Weizenbaum does point to projects in mathematics and chemistry where computers have shown their potential for assisting human scientists in solving problems. He correctly points out that these successes are based on the existence of "strong theories" about their subject matter. We can agree that "common sense" is the human competence hardest to copy in a machine, and that the most constructive advances should come from the wisest division of labor in a synergism of man and his machines.Computers will not give us magical answers to the problems that we, or they, create: with sweat and insight we may be able to develop them as ever more effective tools to serve human needs.

The following review appeared in Creative Computer and in the SIGART News Letter.

AN UNREASONABLE BOOK

Joseph Weizenbaum, *Computer Power and Human Reason*, W. H. Freeman Co., San Francisco 1975

This moralistic and incoherent book uses computer science and technology as an illustration to support the view promoted by Lewis Mumford, Theodore Roszak, and Jacques Ellul, that science has led to an immoral view of man and the world. I am frightened by its arguments that certain research should not be done if it is based on or might result in an "obscene" picture of the world and man. Worse yet, the book's notion of "obscenity" is vague enough to admit arbitrary interpretations by activist bureaucrats.

IT'S HARD TO FIGURE OUT WHAT HE REALLY BELIEVES ...

Weizenbaum's style involves making extreme statements which are later qualified by contradictory statements. Therefore, almost any quotation is out of context, making it difficult to summarize his contentions accurately.

The following passages illustrate the difficulty:

"*In* 1935, *Michael Polanyi*", [British chemist and philosopher of science, was told by] "*Nicolai Bukharin, one of the leading theoreticians of the Russian Communist party, ... [that] 'under socialism the conception of science pursued for its own sake would disappear, for the interests of scientists would spontaneously turn to the problems of the current Five Year Plan.' Polanyi sensed then that 'the scientific outlook appeared to have produced a mechanical conception of man and history in which there was no place for science itself.' And further that 'this conception denied altogether any intrinsic power to thought and thus denied any grounds for claiming freedom of thought.'*" - from page 1. Well, that's clear enough; Weizenbaum favors freedom of thought and science and is worried about threats to them. But on page 265, we have

"*Scientists who continue to prattle on about 'knowledge for its own sake' in order to exploit that slogan for their self-serving ends have detached science and knowledge from any contact with the real world*". Here Weizenbaum seems to be against pure science, i.e. research motivated solely by curiosity. We also have

"*With few exceptions, there have been no results, from over twenty years of artificial intelligence research, that have found their way into industry generally or into the computer industry in particular*". - page 229 This again suggests that industrial results are necessary to validate science.

"*Science promised man power. But as so often happens when people are seduced by promises of power .. the price actually paid is servitude and impotence*". This is from the book jacket. Presumably the publisher regards it as a good summary of the book's main point.

"*I will, in what follows, try to maintain the position that there is nothing wrong with viewing man*

as an information processor (or indeed as anything else) nor with attempting to understand him from that perspective, providing, however, that we never act as though any single perspective can comprehend the whole man." - page 140. We can certainly live with that, but

"*Not only has our unbounded feeding on science caused us to become dependent on it, but, as happens with many other drugs taken in increasing dosages, science has been gradually converted into a slow acting poison*". - page 13. These are qualified by

"*I argue for the rational use of science and technology, not for its mystification, let alone its abandonment*". - page 256

In reference to the proposal for a moratorium on certain experiments with recombinant DNA because they might be dangerous, we have "*Theirs is certainly a step in the right direction, and their initiative is to be applauded. Still, one may ask, why do they feel they have to give a reason for what they recommend at all? Is not the overriding obligation on men, including men of science, to exempt life itself from the madness of treating everything as an object, a sufficient reason, and one that does not even have to be spoken? Why does it have to be explained? It would appear that even the noblest acts of the most well-meaning people are poisoned by the corrosive climate of values of our time.*" Is Weizenbaum against all experimental biology or even all experiments with DNA? I would hesitate to conclude so from this quote; he may say the direct opposite somewhere else. Weizenbaum's goal of getting lines of research abandoned without even having to give a reason seems unlikely to be achieved except in an atmosphere that combines public hysteria and bureaucratic power. This has happened under conditions of religious enthusiasm and in Nazi Germany, in Stalinist Russia and in the China of the "Cultural Revolution". Most likely it won't happen in America.

"*Those who know who and what they are do not need to ask what they should do.*" - page 273. Let me assure the reader that there is nothing in the book that offers any way to interpret this pomposity. I take it as another plea to be free of the bondage of having to give reasons for his denunciations.

The menace of such grandiloquent precepts is that they require a priesthood to apply them to particular cases, and would-be priests quickly crystallize around any potential center of power. A corollary of this is that people can be attacked for what they are rather than for anything specific they have done. The April 1976 issue of *Ms.* has a poignant illustration of this in an article about "trashing".

"*An individual is dehumanized whenever he is treated as less than a whole person*". - page 266. This is also subject to priestly interpretation as in the encounter group movement.

"*The first kind* [of computer application] *I would call simply obscene. These are ones whose very contemplation ought to give rise to feelings of disgust in every civilized person. The proposal I have mentioned, that an animal's visual system and brain be coupled to computers, is an example. It represents an attack on life itself. One must wonder what must have happened to the proposers' perception of life, hence to their perceptions of themselves as part of the continuum of life, that they can even think of such a thing, let alone advocated it*". No argument is offered that might be answered, and no attempt is made to define criteria of acceptability. I think Weizenbaum and the

scientists who have praised the book may be surprised at some of the repressive uses to which the book will be put. However, they will be able to point to passages in the book with quite contrary sentiments, so the repression won't be their fault.

BUT HERE'S A TRY AT SUMMARIZING:

As these inconsistent passages show, it isn't easy to determine Weizenbaum's position, but the following seem to be the book's main points:

1. Computers cannot be made to reason usefully about human affairs.

This is supported by quoting overoptimistic predictions by computer scientists and giving examples of non-verbal human communication. However, Weizenbaum doesn't name any specific task that computers cannot carry out, because he wishes "*to avoid the unnecessary, interminable, and ultimately sterile exercise of making a catalogue of what computers will and will not be able to do, either here and now or ever*". It is also stated that human and machine reasoning are incomparable and that the sensory experience of a human is essential for human reasoning.

2. There are tasks that computers should not be programmed to do.

Some are tasks Weizenbaum thinks shouldn't be done at all - mostly for new left reasons. One may quarrel with his politics, and I do, but obviously computers shouldn't do what shouldn't be done. However, Weizenbaum also objects to computer hookups to animal brains and computer conducted psychiatric interviews. As to the former, I couldn't tell whether he is an anti-vivisectionist, but he seems to have additional reasons for calling them "obscene". The objection to computers doing psychiatric interviews also has a component beyond the conviction that they would necessarily do it badly. Thus he says, "*What can the psychiatrist's image of his patient be when he sees himself, as a therapist, not as an engaged human being acting as a healer, but as an information processor following rules, etc.*?" This seems like the renaissance era religious objections to dissecting the human body that came up when science revived. Even the Popes eventually convinced themselves that regarding the body as a machine for scientific or medical purposes was quite compatible with regarding it as the temple of the soul. Recently they have taken the same view of studying mental mechanisms for scientific or psychiatric purposes.

3. Science has led people to a wrong view of the world and of life.

The view is characterized as mechanistic, and the example of clockwork is given. (It seems strange for a computer scientist to give this example, because the advance of the computer model over older mechanistic models is that computers can and clockwork can't make decisions.) Apparently analysis of a living system as composed of interacting parts rather than treating it as an unanalyzed whole is bad.

4. Science is not the sole or even main source of reliable general knowledge.

However, he doesn't propose any other sources of knowledge or say what the limits of scientific knowledge is except to characterize certain thoughts as "obscene".

5. Certain people and institutions are attacked.

These include the Department of "Defense" (sic), *Psychology Today*, the *New York Times* Data Bank, compulsive computer programmers, Kenneth Colby, Marvin Minsky, Roger Schank, Allen Newell, Herbert Simon, J.W. Forrester, Edward Fredkin, B.F. Skinner, Warren McCulloch (until he was old), Laplace and Leibniz.

6. Certain political and social views are taken for granted.

The view that U.S. policy in Vietnam was "murderous" is used to support an attack on "logicality" (as opposed to "rationality") and the view of science as a "slow acting poison". The phrase "*It may be that the people's cultivated and finally addictive hunger for private automobiles...*" (p.30) makes psychological, sociological, political, and technological presumptions all in one phrase. Similarly, "*Men could instead choose to have truly safe automobiles, decent television, decent housing for everyone, or comfortable, safe, and widely distributed mass transportation.*" presumes wide agreement about what these things are, what is technologically feasible, what the effects of changed policies would be, and what activities aimed at changing people's taste are permissible for governments.

THE ELIZA EXAMPLE

Perhaps the most interesting part of the book is the account of his own program ELIZA that parodies Rogerian non-directive psychotherapy and his anecdotal account of how some people ascribe intelligence and personality to it. In my opinion, it is quite natural for people who don't understand the notion of algorithm to imagine that a computer computes analogously to the way a human reasons. This leads to the idea that accurate computation entails correct reasoning and even to the idea that computer malfunctions are analogous to human neuroses and psychoses. Actually, programming a computer to draw interesting conclusions from premises is very difficult and only limited success has been attained. However, the effect of these natural misconceptions shouldn't be exaggerated; people readily understand the truth when it is explained, especially when it applies to a matter that concerns them. In particular, when an executive excuses a mistake by saying that he placed excessive faith in a computer, a certain skepticism is called for.

Colby's (1973) study is interesting in this connection, but the interpretation below is mine. Colby had psychiatrists interview patients over a teletype line and also had them interview his PARRY program that simulates a paranoid. Other psychiatrists were asked to decide from the transcripts whether the interview was with a man or with a program, and they did no better than chance. However, since PARRY is incapable of the simplest causal reasoning, if you ask, "How do you know the people following you are Mafia" and get a reply that they look like Italians, this must be a man not PARRY. Curiously, it is easier to imitate (well enough to fool a psychiatrist)

the emotional side of a man than his intellectual side. Probably the subjects expected the machine to have more logical ability, and this expectation contributed to their mistakes. Alas, random selection from the directory of the Association for Computing Machinery did no better.

It seems to me that ELIZA and PARRY show only that people, including psychiatrists, often have to draw conclusions on slight evidence, and are therefore easily fooled. If I am right, two sentences of instruction would allow them to do better.

In his 1966 paper on ELIZA (cited as 1965), Weizenbaum writes,

"*One goal for an augmented ELIZA program is thus a system which already has access to a store of information about some aspect of the real world and which, by means of conversational interaction with people, can reveal both what it knows, i.e. behave as an information retrieval system, and where its knowledge ends and needs to be augmented. Hopefully the augmentation of its knowledge will also be a direct consequence of its conversational experience. It is precisely the prospect that such a program will converse with many people and learn something from each of them which leads to the hope that it will prove an interesting and even useful conversational partner.*" Too bad he didn't successfully pursue this goal; no-one else has. I think success would have required a better understanding of formalization than is exhibited in the book.

WHAT DOES HE SAY ABOUT COMPUTERS?

While Weizenbaum's main conclusions concern science in general and are moralistic in character, some of his remarks about computer science and AI are worthy of comment.

1. He concludes that since a computer cannot have the experience of a man, it cannot understand a man. There are three points to be made in reply. First, humans share each other's experiences and those of machines or animals only to a limited extent. In particular, men and women have different experiences. Nevertheless, it is common in literature for a good writer to show greater understanding of the experience of the opposite sex than a poorer writer of that sex. Second, the notion of experience is poorly understood; if we understood it better, we could reason about whether a machine could have a simulated or vicarious experience normally confined to humans. Third, what we mean by understanding is poorly understood, so we don't yet know how to define whether a machine understands something or not.

2. Like his predecessor critics of artificial intelligence, Taube, Dreyfus and Lighthill, Weizenbaum is impatient, implying that if the problem hasn't been solved in twenty years, it is time to give up. Genetics took about a century to go from Mendel to the genetic code for proteins, and still has a long way to go before we will fully understand the genetics and evolution of intelligence and behavior. Artificial intelligence may be just as difficult. My current answer to the question of when machines will reach human-level intelligence is that a precise calculation shows that we are between 1.7 and 3.1 Einsteins and .3 Manhattan Projects away from the goal. However, the current research is producing the information on which the Einstein will base himself and is producing useful capabilities all the time.

3. The book confuses computer simulation of a phenomenon with its formalization in logic. A simulation is only one kind of formalization and not often the most useful - even to a

computer. In the first place, logical and mathematical formalizations can use partial information about a system insufficient for a simulation. Thus the law of conservation of energy tells us much about possible energy conversion systems before we define even one of them. Even when a simulation program is available, other formalizations are necessary even to make good use of the simulation. This review isn't the place for a full explanation of the relations between these concepts.

Like *Punch*'s famous curate's egg, the book is good in parts. Thus it raises the following interesting issues:

1. What would it mean for a computer to hope or be desperate for love? Answers to these questions depend on being able to formalize (not simulate) the phenomena in question. My guess is that adding a notion of hope to an axiomatization of belief and wanting might not be difficult. The study of *propositional attitudes* in philosophical logic points in that direction.

2. Do differences in experience make human and machine intelligence necessarily so different that it is meaningless to ask whether a machine can be more intelligent than a machine? My opinion is that comparison will turn out to be meaningful. After all, most people have not doubt that humans are more intelligent than turkeys. Weizenbaum's examples of the dependence of human intelligence on sensory abilities seem even refutable, because we recognize no fundamental difference in humanness in people who are severely handicapped sensorily, e.g. the deaf, dumb and blind or paraplegics.

IN DEFENSE OF THE UNJUSTLY ATTACKED - SOME OF WHOM ARE INNOCENT

Here are defenses of Weizenbaum's targets. They are not guaranteed to entirely suit the defendees.

Weizenbaum's conjecture that the Defense Department supports speech recognition research in order to be able to snoop on telephone conversations is biased, baseless, false, and seems motivated by political malice. The committee of scientists that proposed the project advanced quite different considerations, and the high officials who made the final decisions are not ogres. Anyway their other responsibilities leave them no time for complicated and devious considerations. I put this one first, because I think the failure of many scientists to defend the Defense Department against attacks they know are unjustified, is unjust in itself, and furthermore has harmed the country.

Weizenbaum doubts that computer speech recognition will have cost-effective applications beyond snooping on phone conversations. He also says, "*There is no question in my mind that there is no pressing human problem that will be more easily solved because such machines exist*". I worry more about whether the programs can be made to work before the sponsor loses patience. Once they work, costs will come down. Winograd pointed out to me that many possible household applications of computers may not be feasible without some computer speech recognition. One needs to think both about how to solve recognized problems and about opportunities to put new technological possibilities to good use. The telephone was not invented by a committee considering already identified problems of communication.

Referring to *Psychology Today* as a cafeteria simply excites the snobbery of those who would like to consider their psychological knowledge to be above the popular level. So far as I know, professional and academic psychologists welcome the opportunity offered by *Psychology Today* to explain their ideas to a wide public. They might even buy a cut-down version of Weizenbaum's book if he asks them nicely. Hmm, they might even buy this review.

Weizenbaum has invented a *New York Times Data Bank* different from the one operated by the *New York Times* - and possibly better. The real one stores abstracts written by humans and doesn't use the tapes intended for typesetting machines. As a result the user has access only to abstracts and cannot search on features of the stories themselves, i.e. he is at the mercy of what the abstractors thought was important at the time.

Using computer programs as psychotherapists, as Colby proposed, would be moral if it would cure people. Unfortunately, computer science isn't up to it, and maybe the psychiatrists aren't either.

I agree with Minsky in criticizing the reluctance of art theorists to develop formal theories. George Birkhoff's formal theory was probably wrong, but he shouldn't have been criticized for trying. The problem seems very difficult to me, and I have made no significant progress in responding to a challenge from Arthur Koestler to tell how a computer program might make or even recognize jokes. Perhaps some reader of this review might have more success.

There is a whole chapter attacking "compulsive computer programmers" or "hackers". This mythical beast lives in the computer laboratory, is an expert on all the ins and outs of the time-sharing system, elaborates the time-sharing system with arcane features that he never documents, and is always changing the system before he even fixes the bugs in the previous version. All these vices exist, but I can't think of any individual who combines them, and people generally outgrow them. As a laboratory director, I have to protect the interests of people who program only part time against tendencies to over-complicate the facilities. People who spend all their time programming and who exchange information by word of mouth sometimes have to be pressed to make proper writeups. The other side of the issue is that we professors of computer science sometimes lose our ability to write actual computer programs through lack of practice and envy younger people who can spend full time in the laboratory. The phenomenon is well known in other sciences and in other human activities.

Weizenbaum attacks the Yale computer linguist, Roger Schank, as follows - the inner quotes are from Schank: "*What is contributed when it is asserted that 'there exists a conceptual base that is interlingual, onto which linguistic structures in a given language map during the understanding process and out of which such structures are created during generation [of linguistic utterances]'? Nothing at all. For the term 'conceptual base' could perfectly well be replaced by the word 'something'. And who could argue with that so-transformed statement?*" Weizenbaum goes on to say that the real scientific problem "remains as untouched as ever". On the next page he says that unless the "Schank-like scheme" understood the sentence "*Will you come to dinner with me this evening?*" to mean "*a shy young man's desperate longing for love*, then the sense in which the system "understands" is "about as weak as the sense in which ELIZA "understood"". This good example raises interesting issues and seems to call for some distinctions. Full understanding of the sentence indeed results in knowing about the young man's desire for love, but it would seem

that there is a useful lesser level of understanding in which the machine would know only that he would like her to come to dinner.

Contrast Weizenbaum's demanding, more-human-than-thou attitude to Schank and Winograd with his respectful and even obsequious attitude to Chomsky. We have "*The linguist's first task is therefore to write grammars, that is, sets of rules, of particular languages, grammars capable of characterizing all and only the grammatically admissible sentences of those languages, and then to postulate principles from which crucial features of all such grammars can be deduced. That set of principles would then constitute a universal grammar. Chomsky's hypothesis is, to put it another way, that the rules of such a universal grammar would constitute a kind of projective description of important aspects of the human mind.*" There is nothing here demanding that the universal grammar take into account the young man's desire for love. As far as I can see, Chomsky is just as much a rationalist as we artificial intelligentsia.

Chomsky's goal of a universal grammar and Schank's goal of a conceptual base are similar, except that Schank's ideas are further developed, and the performance of his students' programs can be compared with reality. I think they will require drastic revision and may not be on the right track at all, but then I am pursuing a rather different line of research concerning how to represent the basic facts that an intelligent being must know about the world. My idea is to start from epistemology rather than from language, regarding their linguistic representation as secondary. This approach has proved difficult, has attracted few practitioners, and has led to few computer programs, but I still think it's right.

Weizenbaum approves of the Chomsky school's haughty attitude towards Schank, Winograd and other AI based language researchers. On page 184, he states, "*many linguists, for example, Noam Chomsky, believe that enough thinking about language remains to be done to occupy them usefully for yet a little while, and that any effort to convert their present theories into computer models would, if attempted by the people best qualified, be a diversion from the main task. And they rightly see no point to spending any of their energies studying the work of the hackers.*"

This brings the chapter on "compulsive computer programmers" alias "hackers" into a sharper focus. Chomsky's latest book *Reflections on Language* makes no reference to the work of Winograd, Schank, Charniak, Wilks, Bobrow or William Woods to name only a few of those who have developed large computer systems that work with natural language and who write papers on the semantics of natural language. The actual young computer programmers who call themselves hackers and who come closest to meeting Weizenbaum's description don't write papers on natural language. So it seems that the hackers whose work need not be studied are Winograd, Schank, et. al. who are professors and senior scientists. The Chomsky school may be embarassed by the fact that it has only recently arrived at the conclusion that the semantics of natural language is more fundamental than its syntax, while AI based researchers have been pursuing this line for fifteen years.

The outside observer should be aware that to some extent this is a pillow fight within M.I.T. Chomsky and Halle are not to be dislodged from M.I.T. and neither is Minsky - whose students have pioneered the AI approach to natural language. Schank is quite secure at Yale. Weizenbaum also has tenure. However, some assistant professorships in linguistics may be at stake, especially at M.I.T.

Allen Newell and Herbert Simon are criticized for being overoptimistic and are considered morally defective for attempting to describe humans as difference-reducing machines. Simon's view that the human is a simple system in a complex environment is singled out for attack. In my opinion, they were overoptimistic, because their GPS model on which they put their bets wasn't good enough. Maybe Newell's current *production system models* will work out better. As to whether human mental structure will eventually turn out to be simple, I vacillate but incline to the view that it will turn out to be one of the most complex biological phenomena.

I regard Forrester's models as incapable of taking into account qualitative changes, and the world models they have built as defective even in their own terms, because they leave out saturation-of-demand effects that cannot be discovered by curve-fitting as long as a sytem is rate-of-expansion limited. Moreover, I don't accept his claim that his models are better suited than the unaided mind in "interpreting how social systems behave", but Weizenbaum's sarcasm on page 246 is unconvincing. He quotes Forrester, "[desirable modes of behavior of the social system] *seem to be possible only if we have a good understanding of the system dynamics and are willing to endure the self-discipline and pressures that must accompany the desirable mode*". Weizenbaum comments, "*There is undoubtedly some interpretation of the words 'system' and 'dynamics' which would lend a benign meaning to this observation*". Sorry, but it looks ok to me provided one is suitably critical of Forrester's proposed social goals and the possibility of making the necessary assumptions and putting them into his models.

Skinner's behaviorism that refuses to assign reality to people's internal state seems wrong to me, but we can't call him immoral for trying to convince us of what he thinks is true.

Weizenbaum quotes Edward Fredkin, former director of Project MAC, and the late Warren McCulloch of M.I.T. without giving their names. pp. 241 and 240. Perhaps he thinks a few puzzles will make the book more interesting, and this is so. Fredkin's plea for research in automatic programming seems to overestimate the extent to which our society currently relies on computers for decisions. It also overestimates the ability of the faculty of a particular university to control the uses to which technology will be put, and it underestimates the difficulty of making knowledge based systems of practical use. Weizenbaum is correct in pointing out that Fredkin doesn't mention the existence of genuine conflicts in society, but only the new left sloganeering elsewhere in the book gives a hint as to what he thinks they are and how he proposes to resolve them.

As for the quotation from (McCulloch 1956), Minsky tells me "this is a brave attempt to find a dignified sense of freedom within the psychological determinism morass". Probably this can be done better now, but Weizenbaum wrongly implies that McCulloch's 1956 effort is to his moral discredit.

Finally, Weizenbaum attributes to me two statements - both from oral presentations - which I cannot verify. One of them is "*The only reason we have not yet succeeded in simulating every aspect of the real world is that we have been lacking a sufficiently powerful logical calculus. I am working on that problem*". This statement doesn't express my present opinion or my opinion in 1973 when I am alleged to have expressed it in a debate, and no-one has been able to find it in the video-tape of the debate.

We can't simulate "every aspect of the real world", because the initial state information is never available, the laws of motion are imperfectly known, and the calculations for a simulation are too extensive. Moreover, simulation wouldn't necessarily answer our questions. Instead, we must find out how to represent in the memory of a computer the information about the real world that is actually available to a machine or organism with given sensory capability, and also how to represent a means of drawing those useful conclusions about the effects of courses of action that can be correctly inferred from the attainable information. Having a *sufficiently powerful logical calculus* is an important part of this problem - but one of the easier parts.

[Note added September 1976 - This statement has been quoted in a large fraction of the reviews of Weizenbaum's book (e.g. in *Datamation* and *Nature*) as an example of the arrogance of the "artificial intelligentsia". Weizenbaum firmly insisted that he heard it in the Lighthill debate and cited his notes as corroboration, but later admitted (in *Datamation*) after reviewing the tape that he didn't, but claimed I must have said it in some other debate. I am confident I didn't say it, because it contradicts views I have held and repeatedly stated since 1959. My present conjecture is that Weizenbaum heard me say something on the importance of formalization, couldn't quite remember what, and quoted "what McCarthy must have said" based on his own misunderstanding of the relation between computer modeling and formalization. (His two chapters on computers show no awareness of the difference between declarative and procedural knowledge or of the discussions in the AI literature of their respective roles). Needless to say, the repeated citation by reviewers of a pompous statement that I never made and which is in opposition to the view that I think represents my major contribution to AI - is very offensive].

The second quotation from me is the rhetorical question, "*What do judges know that we cannot tell a computer*". I'll stand on that if we make it "eventually tell" and especially if we require that it be something that one human can reliably teach another.

A SUMMARY OF POLEMICAL SINS

The speculative sections of the book contain numerous dubious little theories, such as this one about the dehumanizing effect of of the invention of the clock: "*The clock had created literally a new reality; and that is what I meant when I said earlier that the trick man turned that prepared the scene for the rise of modern science was nothing less than the transformation of nature and of his perception of reality. It is important to realize that this newly created reality was and remains an impoverished version of the older one, for it rests on a rejection of those direct experiences that formed the basis for, and indeed constituted the old reality. The feeling of hunger was rejected as a stimulus for eating; instead one ate when an abstract model had achieved a certain state, i.e. when the hand of a clock pointed to certain marks on the clock's face (the anthropomorphism here is highly significant too), and similarly for signals for sleep and rising, and so on.*"

This idealization of primitive life is simply thoughtless. Like modern man, primitive man ate when the food was ready, and primitive man probably had to start preparing it even further in advance. Like modern man, primitive man lived in families whose members are no more likely to become hungry all at once than are the members of a present family.

I get the feeling that in toppling this microtheory I am not playing the game; the theory is intended only to provide an atmosphere, and like the reader of a novel, I am supposed to suspend

disbelief. But the contention that science has driven us from a psychological Garden of Eden depends heavily on such word pictures.

By the way, I recall from my last sabbatical at M.I.T. that the *feeling of hunger* is more often the *direct social stimulus for eating* for the "hackers" deplored in Chapter 4 than it could have been for primitive man. Often on a crisp New England night, even as the clock strikes three, I hear them call to one another, messages flash on the screens, a flock of hackers magically gathers, and the whole picturesque assembly rushes chattering off to Chinatown.

I find the book substandard as a piece of polemical writing in the following respects:

1. The author has failed to work out his own positions on the issues he discusses. Making an extreme statement in one place and a contradictory statement in another is no substitute for trying to take all the factors into account and reach a considered position. Unsuspicious readers can come away with a great variety of views, and the book can be used to support contradictory positions.

2. The computer linguists - Winograd, Schank, et. al. - are denigrated as hackers and compulsive computer programmers by innuendo.

3. One would like to know more precisely what biological and psychological experiments and computer applications he finds acceptable. Reviewers have already drawn a variety of conclusions on this point.

4. The terms "authentic", "obscene", and "dehumanization are used as clubs. This is what mathematicians call "proof by intimidation".

5. The book encourages a snobbery that has no need to argue for its point of view but merely utters code words, on hearing which the audience is supposed applaud or hiss as the case may be. The *New Scientist* reviewer certainly salivates in most of the intended places.

6. Finally, when moralizing is both vehement and vague, it invites authoritarian abuse either by existing authority or by new political movements. Imagine, if you can, that this book were the bible of some bureaucracy, e.g. an Office of Technology Assessment, that acquired power over the computing or scientific activities of a university, state, or country. Suppose Weizenbaum's slogans were combined with *the bureaucratic ethic* that holds that any problem can be solved by a law forbidding something and a bureaucracy of eager young lawyers to enforce it. Postulate further a vague *Humane Research Act* and a "public interest" organization with more eager young lawyers suing to get judges to legislate new interpretations of the Act. One can see a laboratory needing more lawyers than scientists and a Humane Research Administrator capable of forbidding or requiring almost anything.

I see no evidence that Weizenbaum forsees his work being used in this way; he doesn't use the phrase *laissez innover* which is the would-be science bureaucrat's analogue of the economist's *laissez faire*, and he never uses the indefinite phrase "it should be decided" which is a common expression of the bureaucratic ethic. However, he has certainly given his fellow computer scientists at least some reason to worry about potential tyranny.

Let me conclude this section with a quotation from Andrew D. White, the first president of Cornell University, that seems applicable to the present situation - not only in computer science, but also in biology. - "*In all modern history, interference with science in the supposed interest of religion, no matter how conscientious such interference may have been, has resulted in the direst evils both to religion and to science, and invariably; and, on the other hand, all untrammelled scientific investigation, no matter how dangerous to religion some of its stages my have seemed for the time to be, has invariably resulted in the highest good both of religion and of science*". Substitute *morality* for *religion* and the parallel is clear. Frankly, the feebleness of the reaction to attacks on scientific freedom worries me more than the strength of the attacks.

WHAT WORRIES ABOUT COMPUTERS ARE WARRANTED?

Grumbling about Weizenbaum's mistakes and moralizing is not enough. Genuine worries prompted the book, and many people share them. Here are the genuine concerns that I can identify and the opinions of one computer scientist about their resolution: What is the danger that the computer will lead to a false model of man? What is the danger that computers will be misused? Can human-level artificial intelligence be achieved? What, if any, motivational characteristics will it have? Would the achievement of artificial intelligence be good or bad for humanity?

1. Does the computer model lead to a false model of man.

Historically, the mechanistic model of the life and the world followed animistic models in accordance with which, priests and medicine men tried to correct malfunctions of the environment and man by inducing spirits to behave better. Replacing them by mechanistic models replaced shamanism by medicine. Roszak explicitly would like to bring these models back, because he finds them more "human", but he ignores the sad fact that they don't work, because the world isn't constructed that way. The pre-computer mechanistic models of the mind were, in my opinion, unsuccessful,but I think the psychologists pursuing computational models of mental processes may eventually develop a really beneficial psychiatry.

Philosophical and moral thinking hasn't yet found a model of man that relates human beliefs and purposes to the physical world in a plausible way. Some of the unsuccessful attempts have been more mechanistic than others. Both mechanistic and non-mechanistic models have led to great harm when made the basis of political ideology, because they have allowed tortuous reasoning to justify actions that simple human intuition regards as immoral. In my opinion, the relation between beliefs, purposes and wants to the physical world is a complicated but ultimately solvable problem. Computer models can help solve it, and can provide criteria that will enable us to reject false solutions. The latter is more important for now, and computer models are already hastening the decay of dialectical materialism in the Soviet Union.

2. What is the danger that computers will be misused?

Up to now, computers have been just another labor-saving technology. I don't agree with Weizenbaum's acceptance of the claim that our society would have been inundated by paper work

without computers. Without computers, people would work a little harder and get a little less for their work. However, when home terminals become available, social changes of the magnitude of those produced by the telephone and automobile will occur. I have discussed them elsewhere, and I think they will be good – as were the changes produced by the automobile and the telephone. Tyranny comes from control of the police coupled with a tyrannical ideology; data banks will be a minor convenience. No dictatorship yet has been overthrown for lack of a data bank.

One's estimate of whether technology will work out well in the future is correlated with one's view of how it worked out in the past. I think it has worked out well – e.g. cars were not a mistake – and am optimistic about the future. I feel that much current ideology is a combination of older anti-scientific and anti-technological views with new developments in the political technology of instigating and manipulating fears and guilt feelings.

3. What motivations will artificial intelligence have?

It will have what motivations we choose to give it. Those who finally create it should start by motivating it only to answer questions and should have the sense to ask for full pictures of the consequences of alternate actions rather than simply how to achieve a fixed goal, ignoring possible side-effects. Giving it human motivational structure with its shifting goals sensitive to physical state would require a deliberate effort beyond that required to make it behave intelligently.

4. Will artificial intelligence be good or bad?

Here we are talking about machines with the same range of intellectual abilities as are posessed by humans. However, the science fiction vision of robots with almost precisely the ability of a human is quite unlikely, because the next generation of computers or even hooking computers together would produce an intelligence that might be qualitatively like that of a human, but thousands of times faster. What would it be like to be able to put a hundred years thought into every decision? I think it is impossible to say whether qualitatively better answers would be obtained; we will have to try it and see.

The achievement of above-human-level artificial intelligence will open to humanity an incredible variety of options. We cannot now fully envisage what these options will be, but it seems apparent that one of the first uses of high-level artificial intelligence will be to determine the consequences of alternate policies governing its use. I think the most likely variant is that man will use artificial intelligence to transform himself, but once its properties and the conequences of its use are known, we may decide not to use it. Science would then be a sport like mountain climbing; the point would be to discover the facts about the world using some stylized limited means. I wouldn't like that, but once man is confronted by the actuality of full AI, they may find our opinion as relevant to them as we would find the opinion of *Pithecanthropus* about whether subsequent evolution took the right course.

5. What shouldn't computers be programmed to do.

Obviously one shouldn't program computers to do things that shouldn't be done. Moreover, we shouldn't use programs to mislead ourselves or other people. Apart from that, I find none of Weizenbaum's examples convincing. However, I doubt the advisability of making robots with human-like motivational and emotional structures that might have rights and duties independently of humans. Moreover, I think it might be dangerous to make a machine that evolved intelligence by responding to a program of rewards and punishments unless its trainers understand the intellectual and motivational structure being evolved.

All these questions merit and have received more extensive discussion, but I think the only rational policy now is to expect the people confronted by the problem to understand their best interests better than we now can. Even if full AI were to arrive next year, this would be right. Correct decisions will require an intense effort that cannot be mobilized to consider an eventuality that is still remote. Imagine asking the presidential candidates to debate on TV what each of them would do about each of the forms that full AI might take.

References:

McCulloch,W.S.(1956) Toward some circuitry of ethical robots or an observational science of the genesis of social evaluation in the mind-like behavior of artifacts. *Acta Biotheoretica*,XI, *parts* 3/4, 147–156

This review is filed as WEIZEN.REV[PUB,JMC] at SU-AI on the ARPA net. Any comments sent to JMC@SU-AI will be stored in the directory PUB,JMC also known as *McCarthy's Electric Magazine*. The comment files will be designated WEIZEN.1, WEIZEN.2, etc.

John McCarthy
Artificial Intelligence Laboratory
Stanford, California 94305

September 16, 1976

A RESPONSE TO JOHN McCARTHY

by

JOSEPH WEIZENBAUM

Whatever the merit of John McCarthy's review of Computer Power and Human Reason may be, it is all but undone by his repeated assertion that the positions taken in the book are derived from a "new left" political ideology. Not long ago the terms "pinko" or "commie" served the function McCarthy assigns to "new left" in his review. I would have thought, but for this exhibit, that even people with as limited a sense of history as John McCarthy displays in his review might have learned something from the events of the tragic decades the United States has just passed through. I would have thought, but for this exhibit, that all participants in scholarly debates had by now renounced argument by irrelevant political association.

McCarthy's warning to "the outside observer" that the book is motivated by a struggle over academic appointments, tenure decisions, etc. going on within M.I.T. is bizarre and absurd. It is another blot on the review and, I would say, on its author. Such tactics are simply indecent.

I am disturbed by John McCarthy's misreading of my book's "main points." The books actual main point, to the extent that there is one main point, is that no single way of seeing the world, whether it be that of the computer metaphor, of science, of religion, of some political dogma, or of whatever, is sufficient to yield an understanding of the world worthy of the human potential to understand.

McCarthy sees the book making the following main points:

1) COMPUTERS CANNOT BE MADE TO REASON AS POWERFULLY AS HUMANS

I wrote: "...I see no way to put a bound on the degree of intelligence [a computer] could, at least in principle, attain." (p. 210) I then go on to argue that a computer's socialization, that is, its acquisition of knowledge from its experience with the world, must necessarily be different from the socialization of human beings. A computer's intelligence must therefore be always alien to human intelligence with respect to a certain range of human affairs. (p. 213) Nowhere do I limit the computer's "reasoning power." The whole book is, however, an attack on the dogmatic coupling of reason to power. This coupling is so much part of the Zeitgeist that single-mindedly committed technological enthusiasts simply cannot conceive of a discussion of reason -- whether by computers or not -- that is not at the same time centered on questions of power. McCarthy's gratuitous projection of his own preoccupations unto me in the form of his attribution to me of this "main point" is further evidence for that.

2) THINGS AND TASKS THAT COMPUTERS SHOULD NOT BE PROGRAMMED TO DO.

Yes, that is genuinely a main point of the book. And McCarthy is right in observing that task that should not be done at all should not be done by computers either. McCarthy and I agree that psychotherapy should under some circumstances be practiced. I am opposed to machine administered psychotherapy and McCarthy cannot see what objections there might be to it (other than those that arise from "new left" motivations) if it were to "cure" people. Prefontal lobotomy "cures" certain mental disorders, but at what price to the patient and, I would add, to the surgeon as well? I believe that machine administered psychotherapy would induce an image of what it means to be human that would be prohibitively costly to human culture. One may disagree with this belief. But one would first have to understand it and to take it into account.

Elsewhere I say that an individual is dehumanized whenever he is treated as less than a whole person. The relevance of that to the present discussion can be seen if one recalls how inhumanely many surgeons treat their patients and people generally. They have, after many years of seeing their patients mainly as objects to be cut and sewed, come to see them as nothing more than objects. Many surgeons eventually see everyone, most importantly themselves, in this narrow way. Similar remarks apply to other professions. It is of course necessary for all of us to adobt an effectively clinical attitude toward people we deal with in a large variety of situations. The surgeon could not actually cut into living flesh were he not able to impose a psychological distance between himself and his patient while actually wielding the scalpel. But somewhere in his inner being he should hold on to his perception of his patient as a whole person. Even more importantly, the patient must never be led into a situation in which he is forced, or even merely encouraged, to regard himself as a mere object. My fear is that computer administered psychotherapy necessarily induces just this kind of self-image in the patients who would be subject to it. That, basically, is my objection to it. I cannot see how such a system could "cure" people in any reasonable sense of the word "cure", that is, in a sufficiently encompassing interpretation of that word.

3) SCIENCE HAS LED PEOPLE TO A WRONG VIEW OF THE WORLD AND LIFE.

There is no "correct" or "wrong" view of life to which science or anything else can lead. The point McCarthy here misconstrues is that science, or any other system of thought, leads to an impoverished view of the world and of life when it or any system is taken to be the only legitimate perspective on the world and on life.

4) SCIENCE IS NOT THE SOLE OR EVEN THE MAIN SOURCE OF KNOWLEDGE.

How reliable would McCarthy say is his knowledge that his children are biologically his children or that the person he knows as his father is his biological father? Is science his source of such knowledge? What proportion of the truly important actions McCarthy has taken in the course of his adult life were predicated solely or even mainly on knowledge he validated by appeals to science? Did the ancients have reliable knowledge? Or have we had reliable knowledge only since the founding of the British Royal Society -- or that of the Stanford AI Lab -- or not yet at all?

5) CERTAIN PEOPLE AND INSTITUTIONS ARE BAD.

I know of very few people I would call "bad" -- Hitler and Himmler are examples. I think some of the people McCarthy lists are often wrong and sometimes behave irresponsibly especially when they speak to and write for lay audiences. There appears to be wide agreement on that within the AI community itself. I think the views expressed by some of the people mentioned are dangerous. These views should be discussed, not suppressed. My book contributes to the required discussion. Some people will surely find my views wrong and perhaps even dangerous. If they think my views worthy of more than contempt, they should discuss them. Does McCarthy think I am "bad?" I don't believe so. Why then should he believe I think the people he mentions are bad?

The Department of "Defense" is, in my view, on the whole bad. And the quotation marks are, I would guess, entirely appropriate in the eyes of most of the people of the world -- especially in the eyes of many who have read Orwell. If the emperor wears no clothes, we should say he wears no clothes.

Other remarks:

I do not say and I do not believe that "if the problem hasn't been solved in twenty years, we should give up." I say (p. 198) "... it would be wrong ... to make impossibility arguments about what computers can do entirely on the grounds of our present ignorance." That is quite the opposite of what McCarthy charges me with saying.

I do not say or imply that "the Defense Department supports speech recognition research in order to be able to snoop on telephone conversations." Attributing this view to me is, in McCarthy's words, "biased, baseless, false, and [seemingly] motivated by malice." I wrote: "This project then represents, in the eyes of its chief sponsor, a long step toward a fully automated battlefield." I then state my opinion

that, should we get speech recognition, large organizations such as the government would use it for snooping, etc. I believe that my belief is buttressed by the revelations (N.Y.T., August 31 1975) that the "NSA eavesdrops on virtually all cable, Telex and other non-telephone communications leaving the U.S. and uses computers to sort and obtain intelligence from the contents ..." (emphasis mine). Clearly the exclusion of telephone communications from this operation is a consequence of only technical limitations that would be removed if we had automatic speech recognition systems.

The reference cited in note 9, page 266, is anonymous because the person in question granted me permission to quote from his internal memorandum on the condition that I not cite his name. It is not nice of John McCarthy to press me to violate my word.

I do not "idealize the life of primitive man." It is a cheap shot often practiced by technological enthusiasts to charge anyone who mentions a loss entailed by man's commitment to technology (and there surely have been and are losses) with advocating a return to pretechnological times. Every modern writer I know of knows that there cannot have been a pretechnological time in the history of what we would call man, and that history cannot be reversed. But it is important that we recognize and understand the costs associated with our current commitment to technology and that we seek for ways to reduce the costs we deem too high. There is nothing anti-technological, anti-scientific, or anti-intellectual in that.

McCarthy suggests that my statement "Those who know who and what they are do not need to ask what they should do" is "menacing" in that he believes it to require a priesthood to apply it to a particular case. The statement appears in a context (p. 273) in which I had just alluded to the fact that people are constantly asking experts what they should do. I don't believe people need "expert" guidance on moral questions. The statement, on its very face, argues that no priesthood is ever necessary to tell people what they must do. I find McCarthy's exact opposite analysis extremely puzzling.

My assertion that "An individual is dehumanized whenever he is treated as less than a whole person" is simply a statement of fact. It is incomprehensible to me that shame or guilt fall on me because McCarthy believes similar statements to be part of the catechism of the "encounter group movement," about which, by the way, I know next to nothing.

Does John McCarthy have a logical calculus within which he has proved that any idea held by the new left or the encounter group movement or by Mumford, Roszak, or Ellul is wrong, a menace, and certain to be used as part of the arsenal of "priests [who] quickly crystallize around any potential center of power?" (Here again we see evidence of McCarthy's preoccupation with power.)

Finally, McCarthy asserts "Philosophical and moral thinking has never found a model of man that relates human beliefs to the physical world in a plausible way." Only someone who has mastered the entire philosophical and moral literature could have the authority to say that. What truly God-like humility! The distance that separates John McCarthy from Joseph Weizenbaum is truly measured by the challenges these two hurl at one another: McCarthy defies Weizenbaum to "Show me a way to knowledge besides science!" And Weizenbaum responds: "Can there be a way toward an authentic model of man that does not include and ultimately rest on philosophical and moral thinking?"

No wonder we talk past one another.
